Department of Statistics
MALAYSIA

BANCI MALAYSIA

POPULATION & HOUSING CENSUS MALAYSIA 2020
YOUR DATA IS OUR FUTURE

#MyCensus2020

Make Sure You're COUNTED

PRICE INTELLIGENCE (PI) – DATA GATHERING BY UTILIZING WEB CRAWLING

Big Data Applications and Utilising Non-Traditional Data Sources and Methods for Official Statistics

10 June 2021

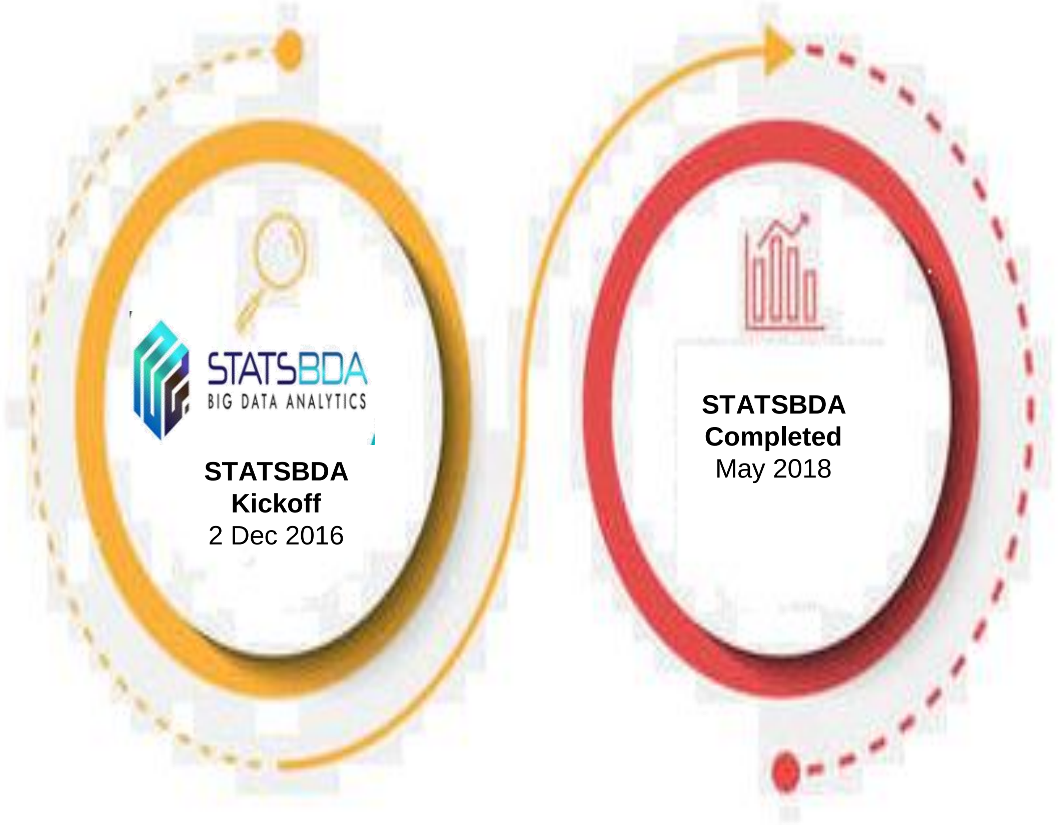
Ms. Mazliana Mustapa
Department of Statistics Malaysia

MYSTATS DAY
20 OCT

SUSTAINABLE DEVELOPMENT GOALS MALAYSIA
2016 - 2030

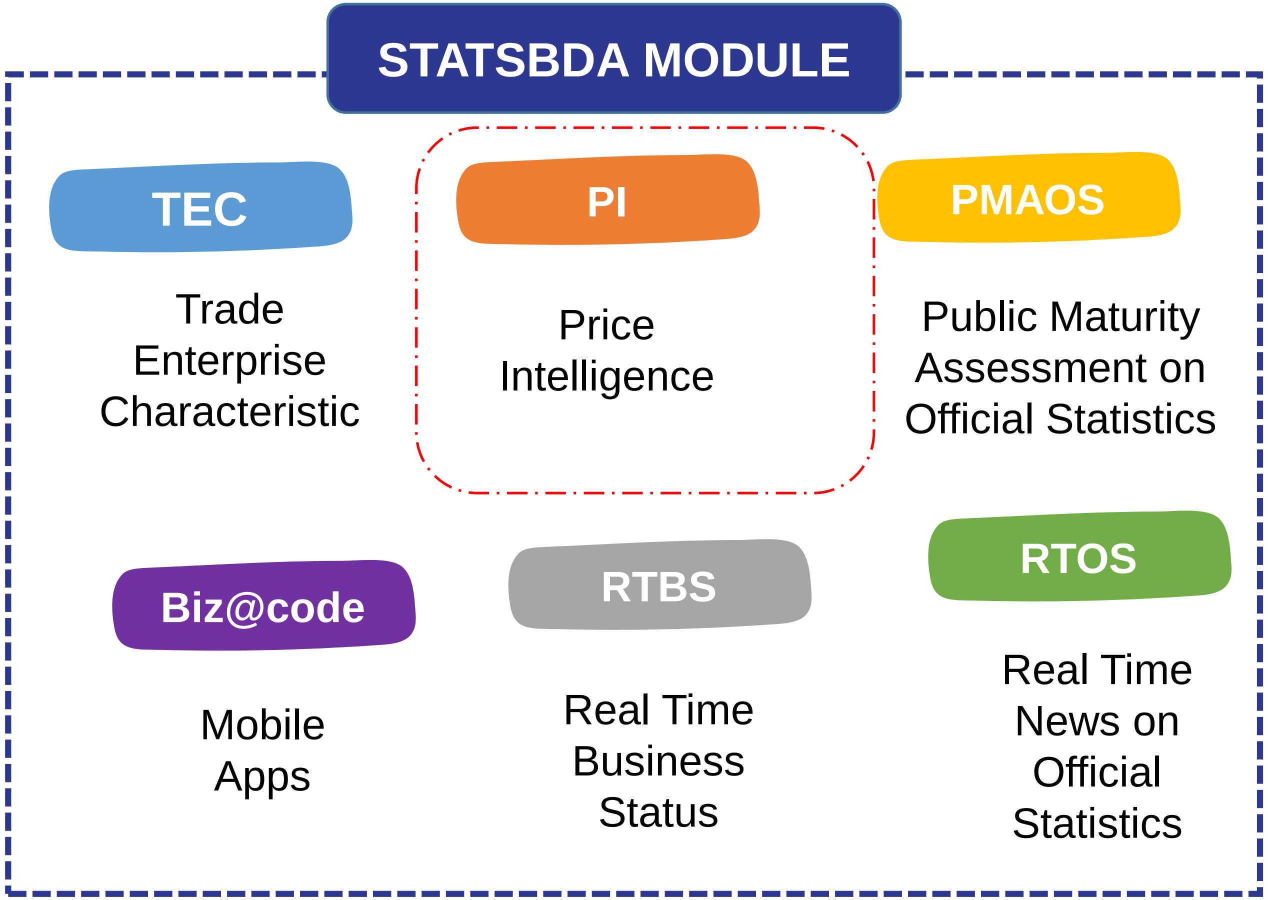
STATSBDA

The Department of Statistics Malaysia (DOSM) has initiated the implementation of BDA under the project of Statistics Big Data Analytics (STATSBDA).



STATSBDA Kickoff
2 Dec 2016

STATSBDA Completed
May 2018



STATSBDA MODULE

TEC
Trade Enterprise Characteristic

PI
Price Intelligence

PMAOS
Public Maturity Assessment on Official Statistics

Biz@code
Mobile Apps

RTBS
Real Time Business Status

RTOS
Real Time News on Official Statistics

STATSBDA OBJECTIVE

Objective

To reduce respondents burden

To use as supplements for existing data in production of certain statistics

To allow high accuracy

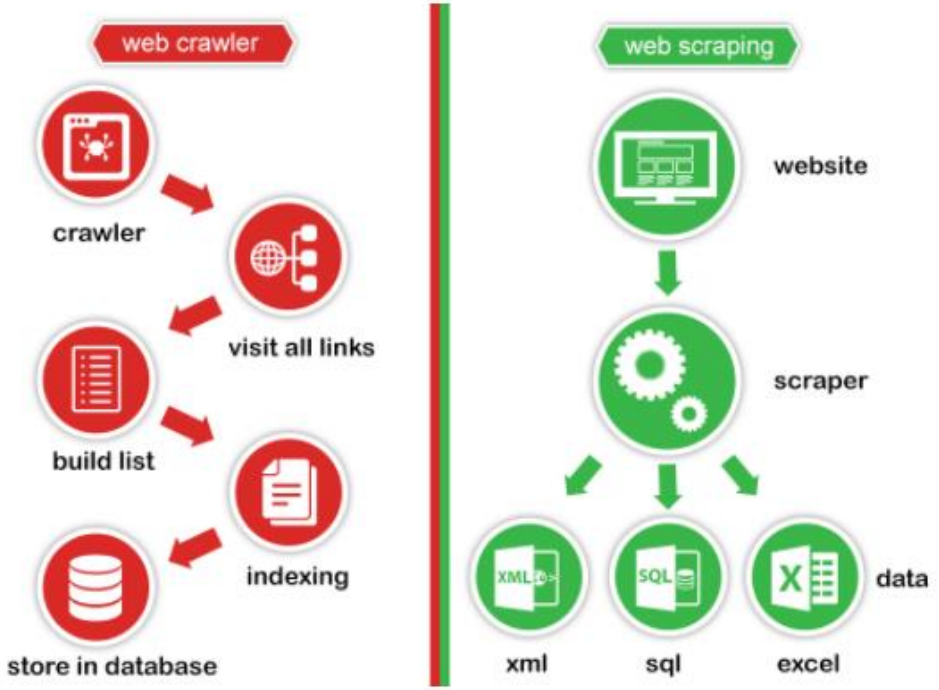
To modernise data collection

To produce new statistical indicators

PRICE INTELLIGENCE (PI)

Price Intelligence

- leveraging the capability of Big Data in collecting large data from various sources and transform them into better structure
- different prices of the same good can be obtained through various online retailer websites, providing a modernized price data collection
- Transform from unstructured data into structured data to perform analysis



What is web crawler?

process of repetitively finding and fetching hyperlinks starting from a list of starting URLs.

What is web scraping?

Web-scraping is automatically retrieving and processing information from websites

PRICE INTELLIGENCE OBJECTIVE

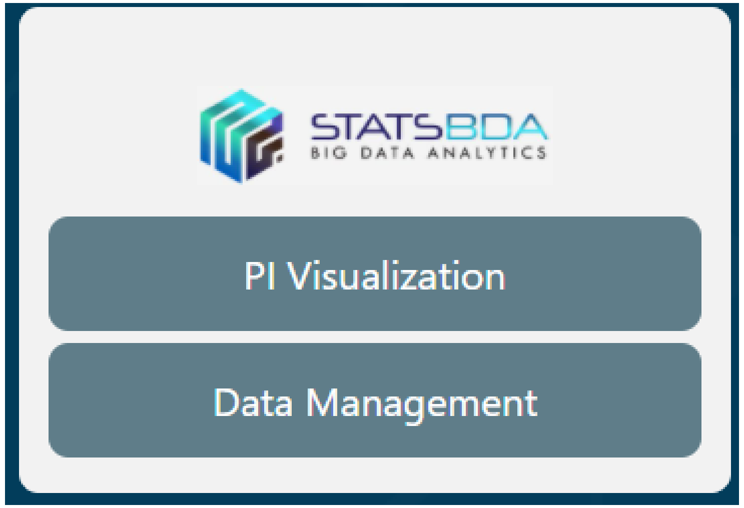
Objective

01

to give better insight in consumer price analysis and monitoring

02

to establish new price basket analysis which will be used as value added to the current Consumer Price Index

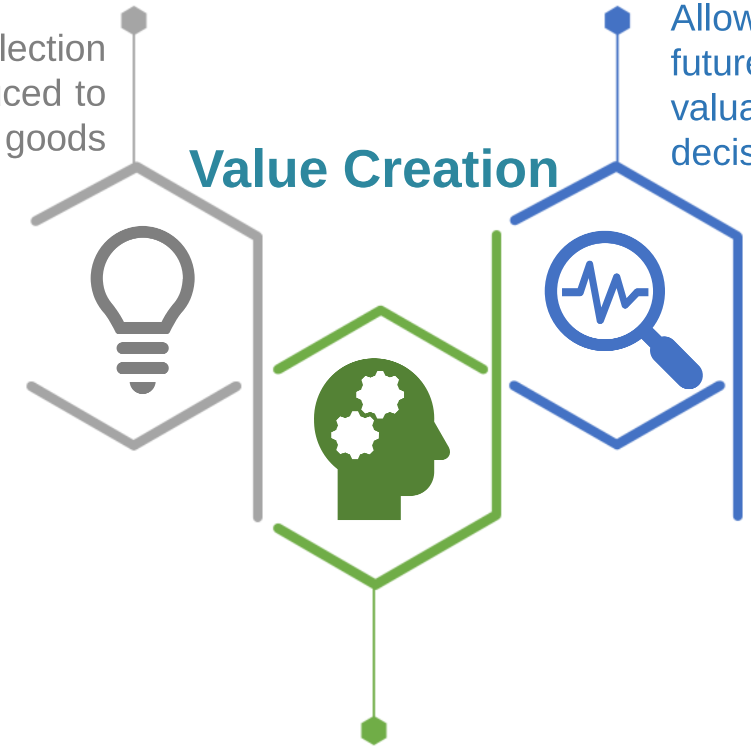


STATSBDA BIG DATA ANALYTICS

New Data Collection Methodology

New data collection process introduced to cover price of goods selling online

Value Creation



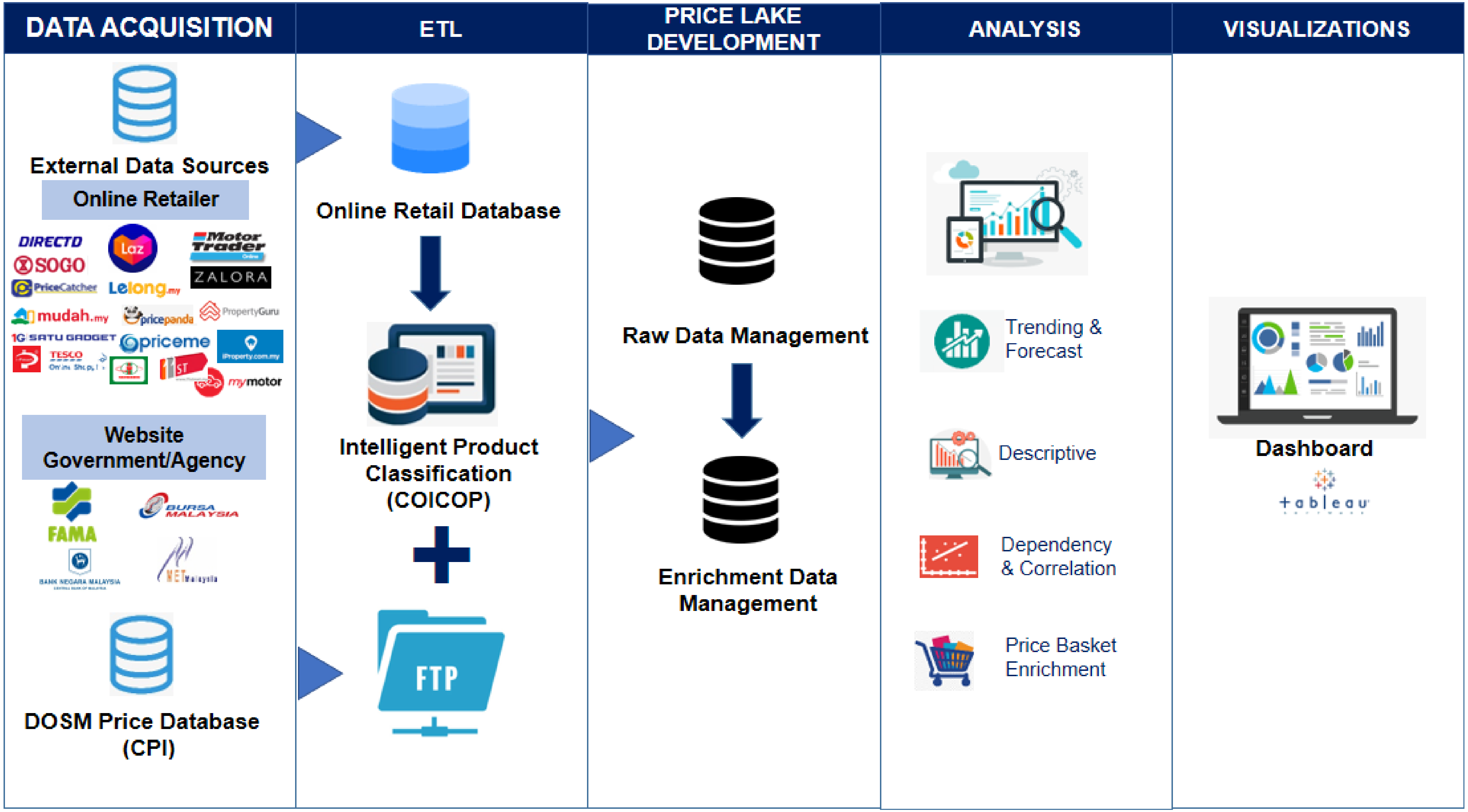
Holistic View in Online and Offline Prices

Allow monitoring and forecast future price trend and as valuable input for price control decisions by government

Transform of Business Process

New data collected enable to create holistic landscape of CPI monitoring process

PRICE INTELLIGENCE ARCHITECTURE



DATA ACQUISITION

External Data Sources

Online Retailer

Website Government/Agency

DOSM Price Database (CPI)

ETL

Online Retail Database

Intelligent Product Classification (COICOP)

PRICE LAKE DEVELOPMENT

Raw Data Management

Enrichment Data Management

ANALYSIS

Trending & Forecast

Descriptive

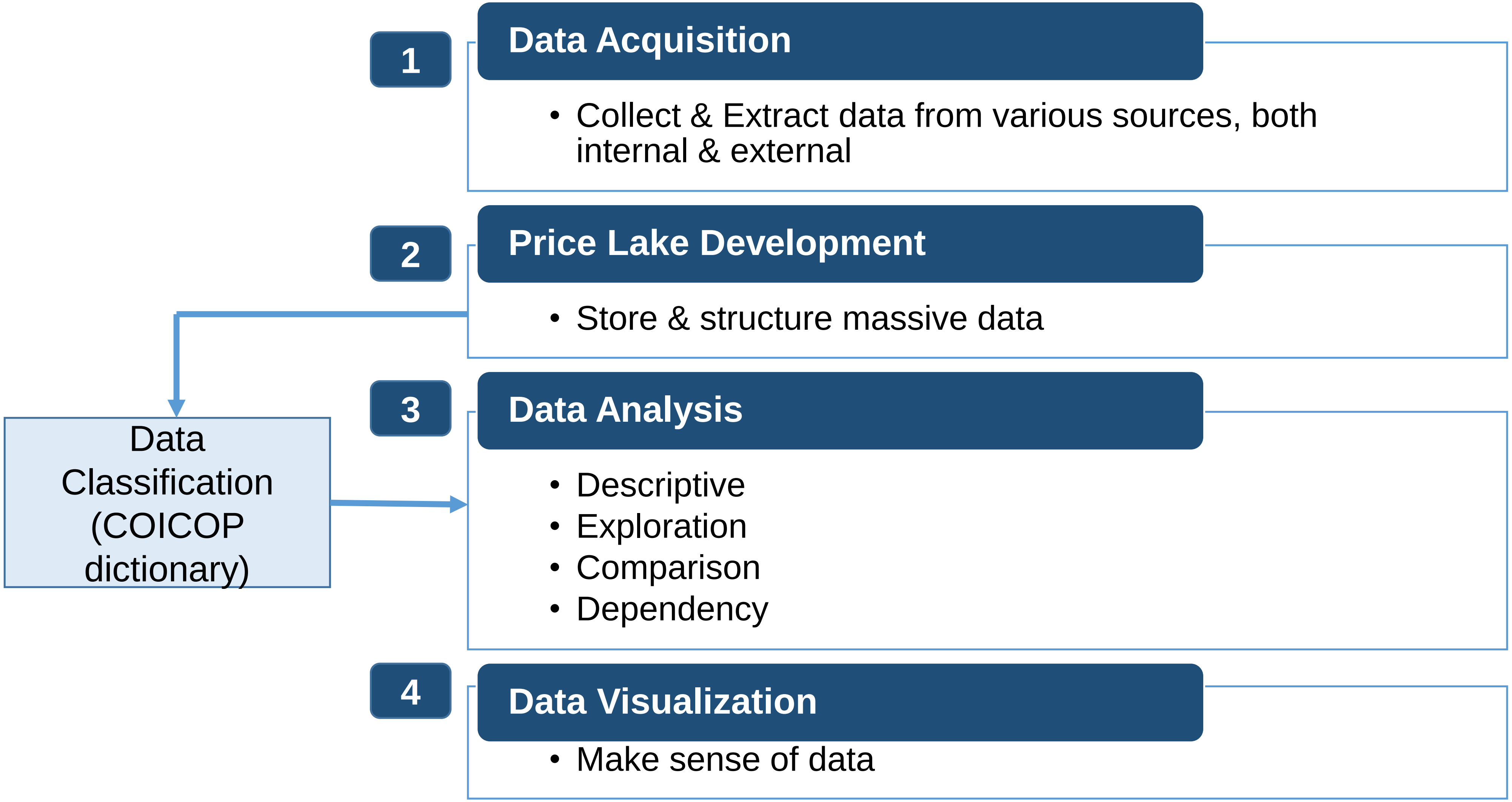
Dependency & Correlation

Price Basket Enrichment

VISUALIZATIONS

Dashboard

MODULE IN PRICE INTELLIGENCE



1 **Data Acquisition**

- Collect & Extract data from various sources, both internal & external

2 **Price Lake Development**

- Store & structure massive data

3 **Data Analysis**

- Descriptive
- Exploration
- Comparison
- Dependency

4 **Data Visualization**

- Make sense of data

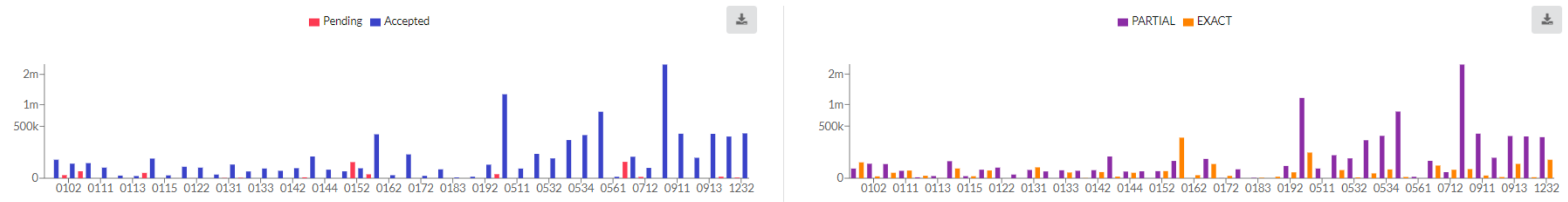
Data Classification (COICOP dictionary)

Note : COICOP → Classification of Individual Consumption According to Purpose

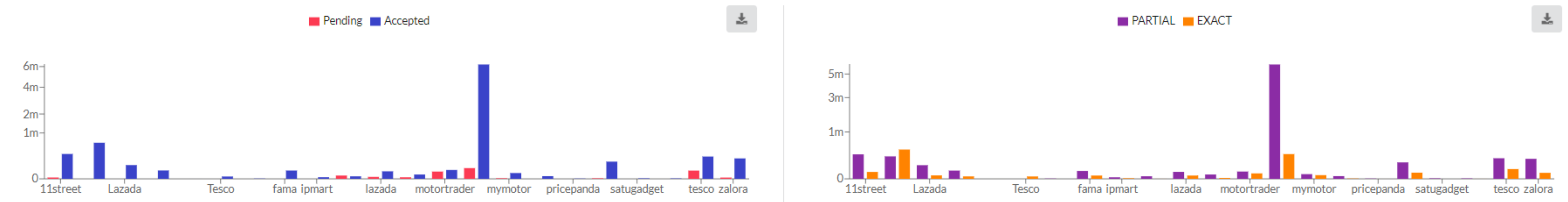
DATA MANAGEMENT

PI Dictionary

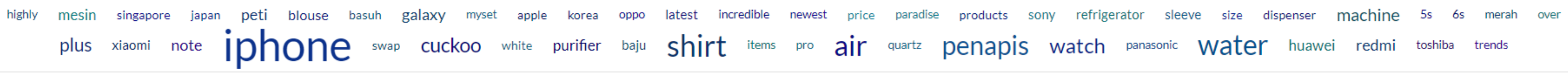
By 4D Code



By Online Source



The word most often appear on the item description



Pending/Accepted refers to the matching result at the 4 Digit COICOP level

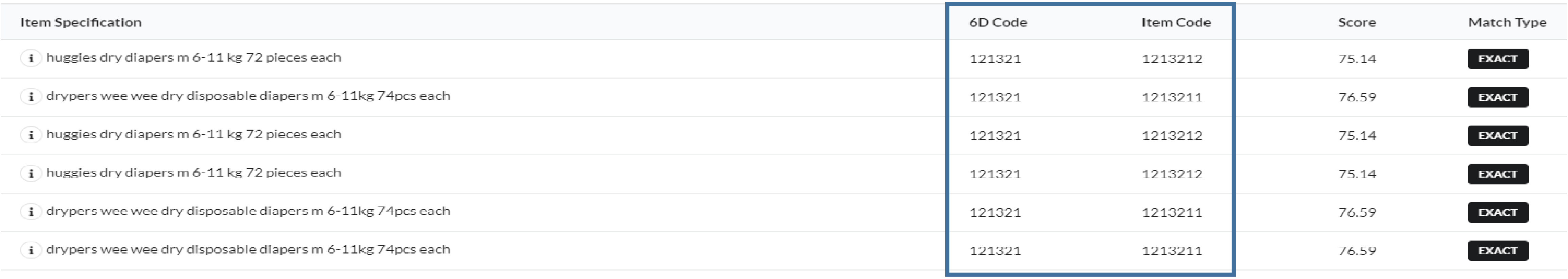
Partially/exact refers to the matching result with MCOICOP

11.3% exact match
88.7% partially match

*Information based on 231 item specifications updates in PI Dictionary

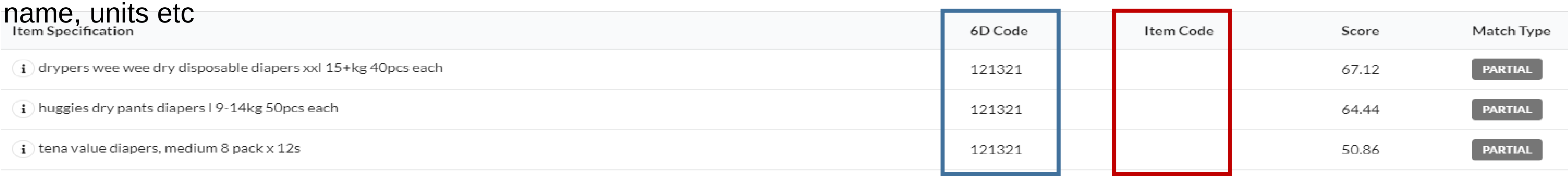
PI Dictionary – Text matching scoring

Specification that meet requirement of items in CPI basket of goods with high scoring value will return exact match, mapped to 7 digits item specification MCOICOP code



Item Specification	6D Code	Item Code	Score	Match Type
huggies dry diapers m 6-11 kg 72 pieces each	121321	1213212	75.14	EXACT
drypers wee wee dry disposable diapers m 6-11kg 74pcs each	121321	1213211	76.59	EXACT
huggies dry diapers m 6-11 kg 72 pieces each	121321	1213212	75.14	EXACT
huggies dry diapers m 6-11 kg 72 pieces each	121321	1213212	75.14	EXACT
drypers wee wee dry disposable diapers m 6-11kg 74pcs each	121321	1213211	76.59	EXACT
drypers wee wee dry disposable diapers m 6-11kg 74pcs each	121321	1213211	76.59	EXACT

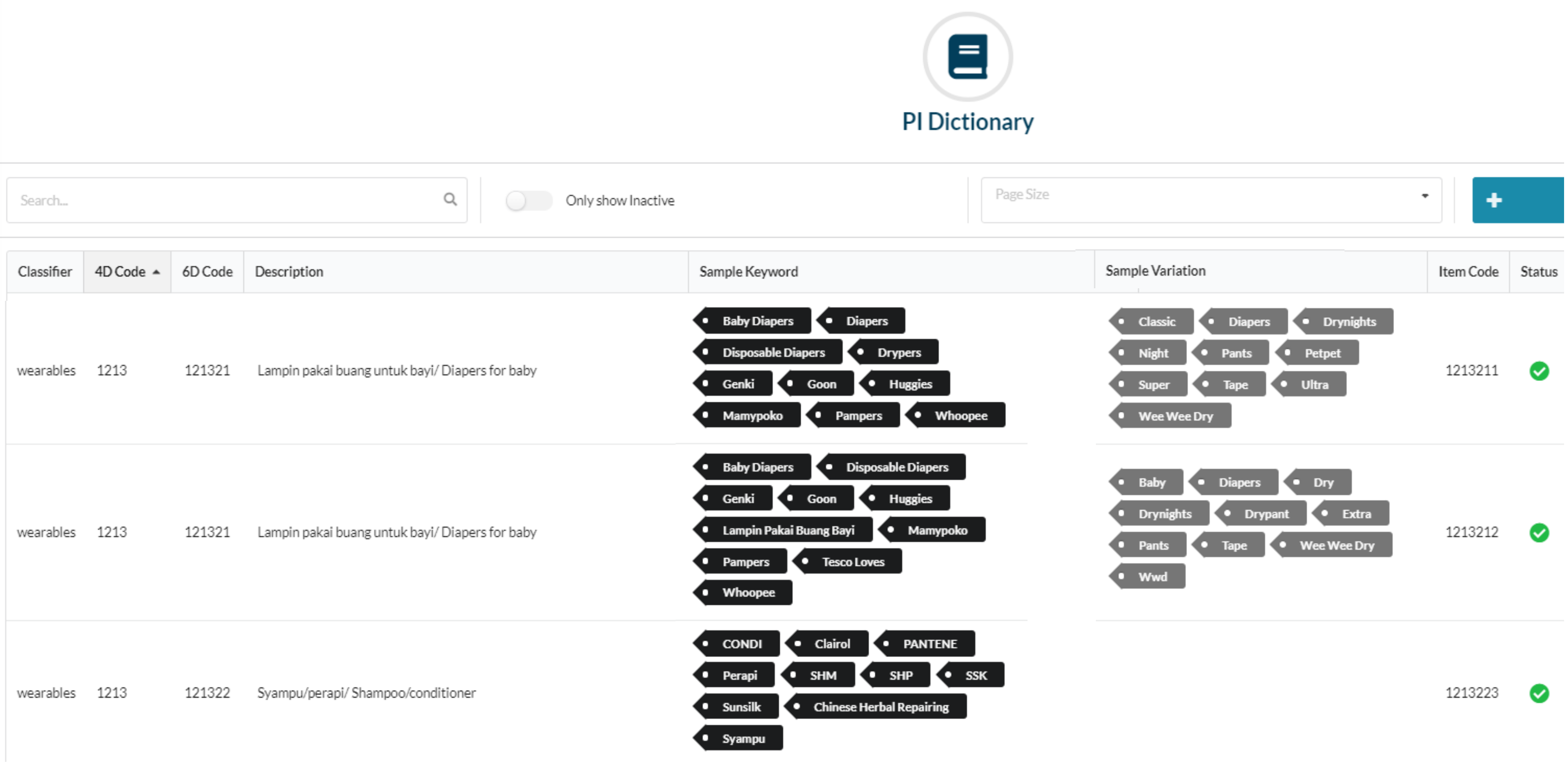
Meanwhile, broad specification items has only 6 digits item code to be matched with online item specification. Items is considered matching as long as the item is the same regardless of brand name, units etc



Item Specification	6D Code	Item Code	Score	Match Type
drypers wee wee dry disposable diapers xxl 15+kg 40pcs each	121321		67.12	PARTIAL
huggies dry pants diapers l 9-14kg 50pcs each	121321		64.44	PARTIAL
tena value diapers, medium 8 pack x 12s	121321		50.86	PARTIAL

DATA MANAGEMENT

PI Dictionary



PI Dictionary

Classifier	4D Code ▲	6D Code	Description	Sample Keyword	Sample Variation	Item Code	Status
wearables	1213	121321	Lampin pakai buang untuk bayi/ Diapers for baby	Baby Diapers Diapers Disposable Diapers Drypers Genki Goon Huggies Mamypoko Pampers Whoopee	Classic Diapers Drynights Night Pants Petpet Super Tape Ultra Wee Wee Dry	1213211	✓
wearables	1213	121321	Lampin pakai buang untuk bayi/ Diapers for baby	Baby Diapers Disposable Diapers Genki Goon Huggies Lampin Pakai Buang Bayi Mamypoko Pampers Tesco Loves Whoopee	Baby Diapers Dry Drynights Drypant Extra Pants Tape Wee Wee Dry Wwd	1213212	✓
wearables	1213	121322	Syampu/perapi/ Shampoo/conditioner	CONDI Clairol PANTENE Perapi SHM SHP SSK Sunsilk Chinese Herbal Repairing Syampu		1213223	✓

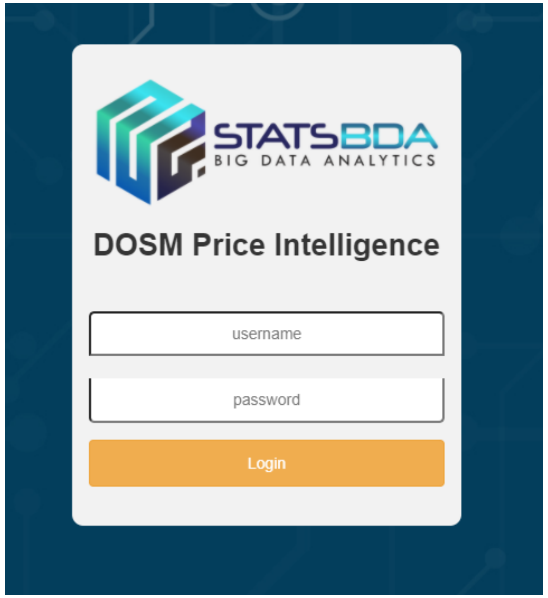
PI Dictionary has been build in order to get the best match to MCOICOP 6/7 digits based on description of the items using text matching scoring

Dictionary will be updated regularly based on basket of goods in CPI

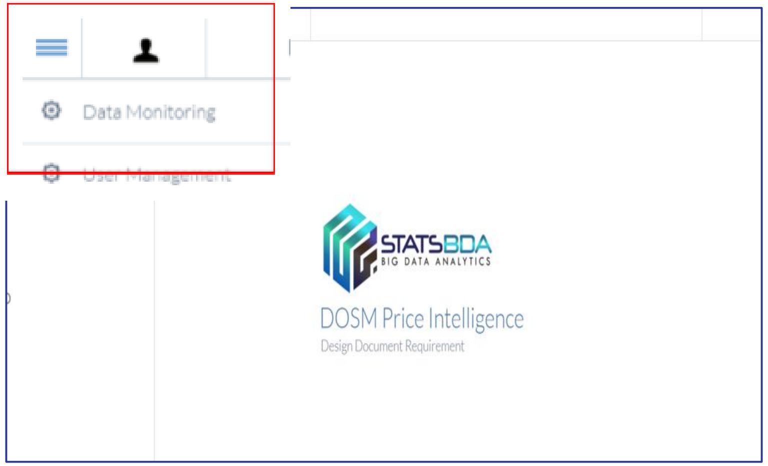
COICOP : Classification of Individual Consumption According to Purpose

DATA SET GENERATOR

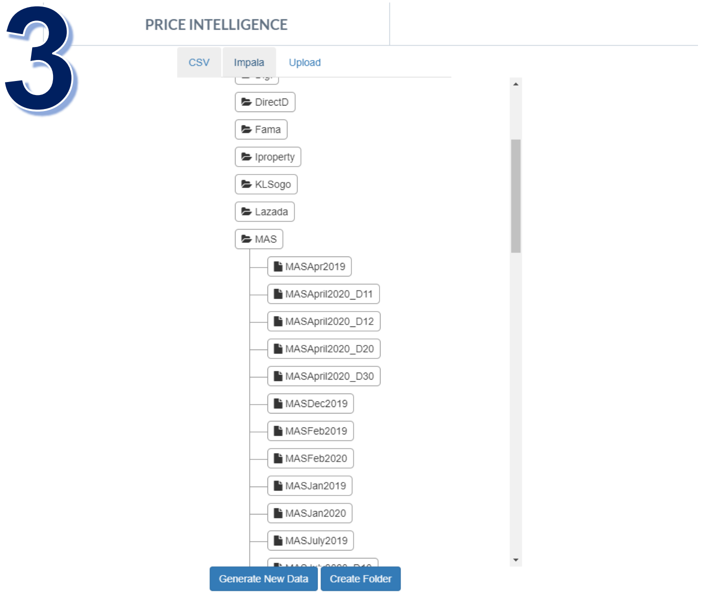
1



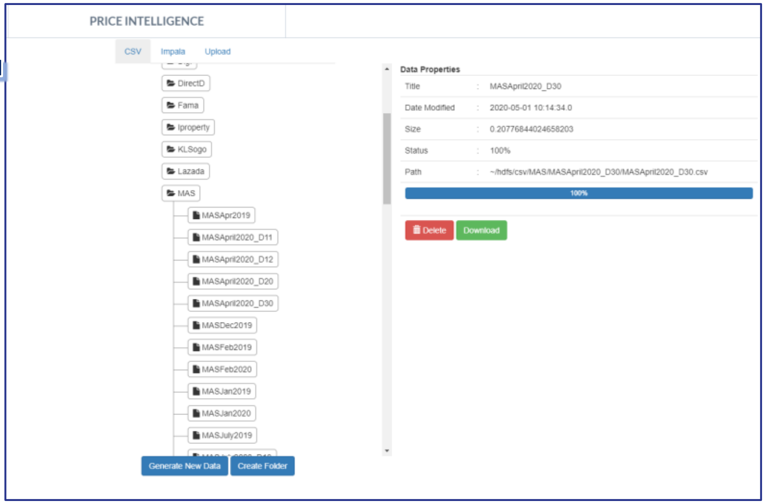
2



3



4

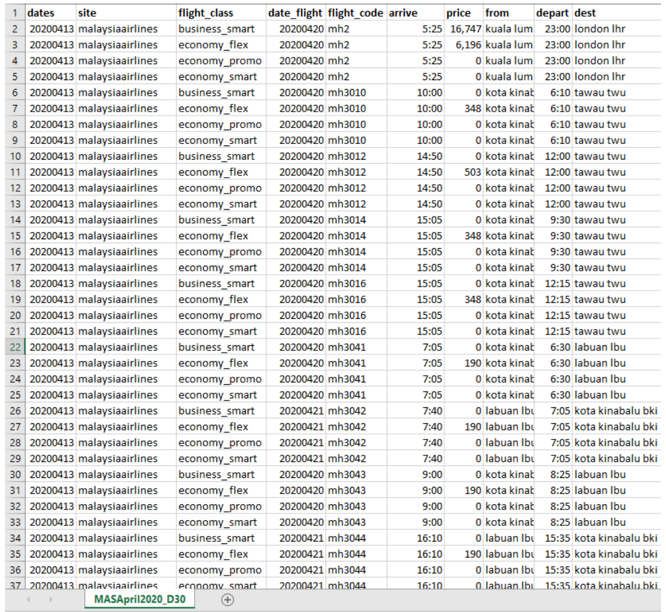


Data Properties

Title : MASApril2020_D30
Date Modified : 2020-05-01 10:14:34.0
Size : 0.20776844024658203
Status : 100%
Path : ~/hdfs/csv/MAS/MASApril2020_D30/MASApril2020_D30.csv

100%

5

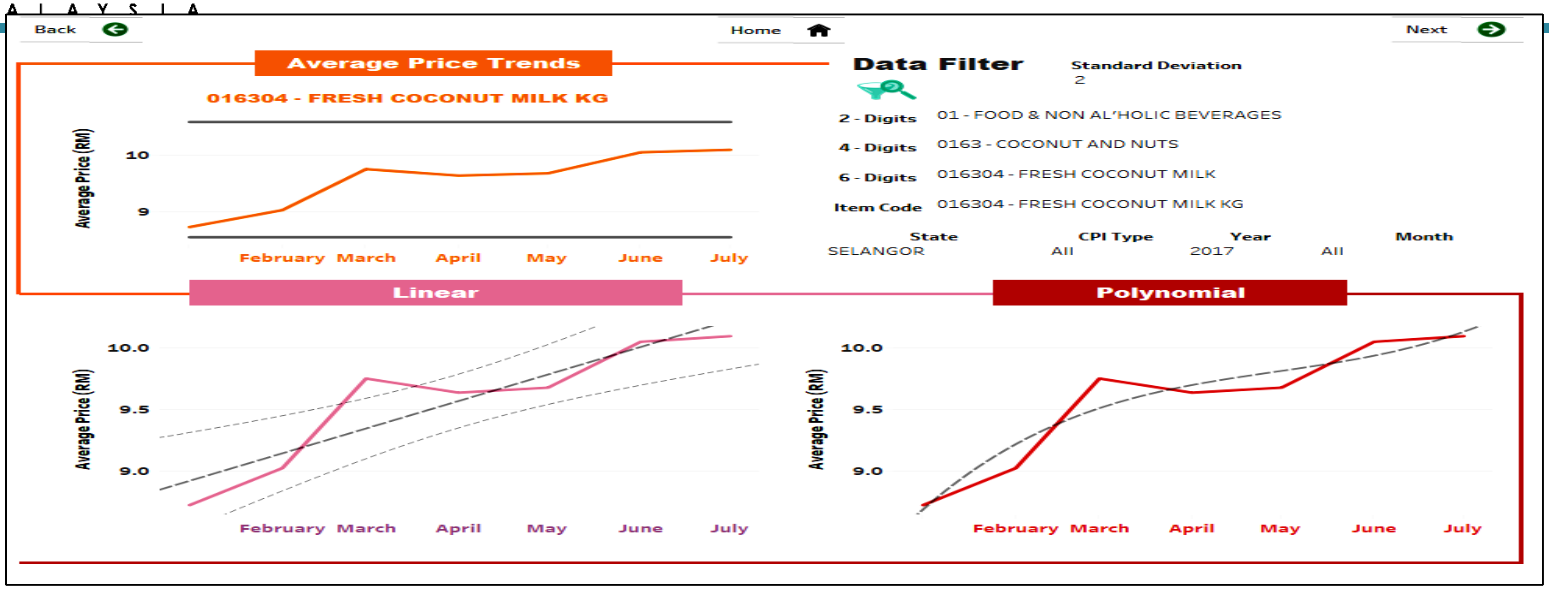


1	dates	site	flight_class	date_flight	flight_code	arrive	price	from	depart	dest
2	20200413	malaysiaairlines	business_smart	20200420	mh2	5:25	16,747	kuala lum	23:00	london lhr
3	20200413	malaysiaairlines	economy_flex	20200420	mh2	5:25	6,196	kuala lum	23:00	london lhr
4	20200413	malaysiaairlines	economy_promo	20200420	mh2	5:25	0	kuala lum	23:00	london lhr
5	20200413	malaysiaairlines	economy_smart	20200420	mh2	5:25	0	kuala lum	23:00	london lhr
6	20200413	malaysiaairlines	business_smart	20200420	mh3010	10:00	0	kota kinak	6:10	tawau twu
7	20200413	malaysiaairlines	economy_flex	20200420	mh3010	10:00	348	kota kinak	6:10	tawau twu
8	20200413	malaysiaairlines	economy_promo	20200420	mh3010	10:00	0	kota kinak	6:10	tawau twu
9	20200413	malaysiaairlines	economy_smart	20200420	mh3010	10:00	0	kota kinak	6:10	tawau twu
10	20200413	malaysiaairlines	business_smart	20200420	mh3012	14:50	0	kota kinak	12:00	tawau twu
11	20200413	malaysiaairlines	economy_flex	20200420	mh3012	14:50	503	kota kinak	12:00	tawau twu
12	20200413	malaysiaairlines	economy_promo	20200420	mh3012	14:50	0	kota kinak	12:00	tawau twu
13	20200413	malaysiaairlines	economy_smart	20200420	mh3012	14:50	0	kota kinak	12:00	tawau twu
14	20200413	malaysiaairlines	business_smart	20200420	mh3014	15:05	0	kota kinak	9:30	tawau twu
15	20200413	malaysiaairlines	economy_flex	20200420	mh3014	15:05	348	kota kinak	9:30	tawau twu
16	20200413	malaysiaairlines	economy_promo	20200420	mh3014	15:05	0	kota kinak	9:30	tawau twu
17	20200413	malaysiaairlines	economy_smart	20200420	mh3014	15:05	0	kota kinak	9:30	tawau twu
18	20200413	malaysiaairlines	business_smart	20200420	mh3016	15:05	0	kota kinak	12:15	tawau twu
19	20200413	malaysiaairlines	economy_flex	20200420	mh3016	15:05	348	kota kinak	12:15	tawau twu
20	20200413	malaysiaairlines	economy_promo	20200420	mh3016	15:05	0	kota kinak	12:15	tawau twu
21	20200413	malaysiaairlines	economy_smart	20200420	mh3016	15:05	0	kota kinak	12:15	tawau twu
22	20200413	malaysiaairlines	business_smart	20200420	mh3041	7:05	0	kota kinak	6:30	labuan lbu
23	20200413	malaysiaairlines	economy_flex	20200420	mh3041	7:05	190	kota kinak	6:30	labuan lbu
24	20200413	malaysiaairlines	economy_promo	20200420	mh3041	7:05	0	kota kinak	6:30	labuan lbu
25	20200413	malaysiaairlines	economy_smart	20200420	mh3041	7:05	0	kota kinak	6:30	labuan lbu
26	20200413	malaysiaairlines	business_smart	20200421	mh3042	7:40	0	labuan lbu	7:05	kota kinabalu bki
27	20200413	malaysiaairlines	economy_flex	20200421	mh3042	7:40	190	labuan lbu	7:05	kota kinabalu bki
28	20200413	malaysiaairlines	economy_promo	20200421	mh3042	7:40	0	labuan lbu	7:05	kota kinabalu bki
29	20200413	malaysiaairlines	economy_smart	20200421	mh3042	7:40	0	labuan lbu	7:05	kota kinabalu bki
30	20200413	malaysiaairlines	business_smart	20200420	mh3043	9:00	0	kota kinak	8:25	labuan lbu
31	20200413	malaysiaairlines	economy_flex	20200420	mh3043	9:00	190	kota kinak	8:25	labuan lbu
32	20200413	malaysiaairlines	economy_promo	20200420	mh3043	9:00	0	kota kinak	8:25	labuan lbu
33	20200413	malaysiaairlines	economy_smart	20200420	mh3043	9:00	0	kota kinak	8:25	labuan lbu
34	20200413	malaysiaairlines	business_smart	20200421	mh3044	16:10	0	labuan lbu	15:35	kota kinabalu bki
35	20200413	malaysiaairlines	economy_flex	20200421	mh3044	16:10	190	labuan lbu	15:35	kota kinabalu bki
36	20200413	malaysiaairlines	economy_promo	20200421	mh3044	16:10	0	labuan lbu	15:35	kota kinabalu bki
37	20200413	malaysiaairlines	economy_smart	20200421	mh3044	16:10	0	labuan lbu	15:35	kota kinabalu bki

Download Data

1. Click Data Monitoring on the top left of the screen
2. Click Folder Name and Click file name
3. Status of the generated data will be displayed.
   Click Download to download to local PC

PRICE INTELLIGENCE ANALYTICS



Data Filter

Standard Deviation
2

2 - Digits 01 - FOOD & NON AL'HOLIC BEVERAGES
4 - Digits 0163 - COCONUT AND NUTS
6 - Digits 016304 - FRESH COCONUT MILK
Item Code 016304 - FRESH COCONUT MILK KG

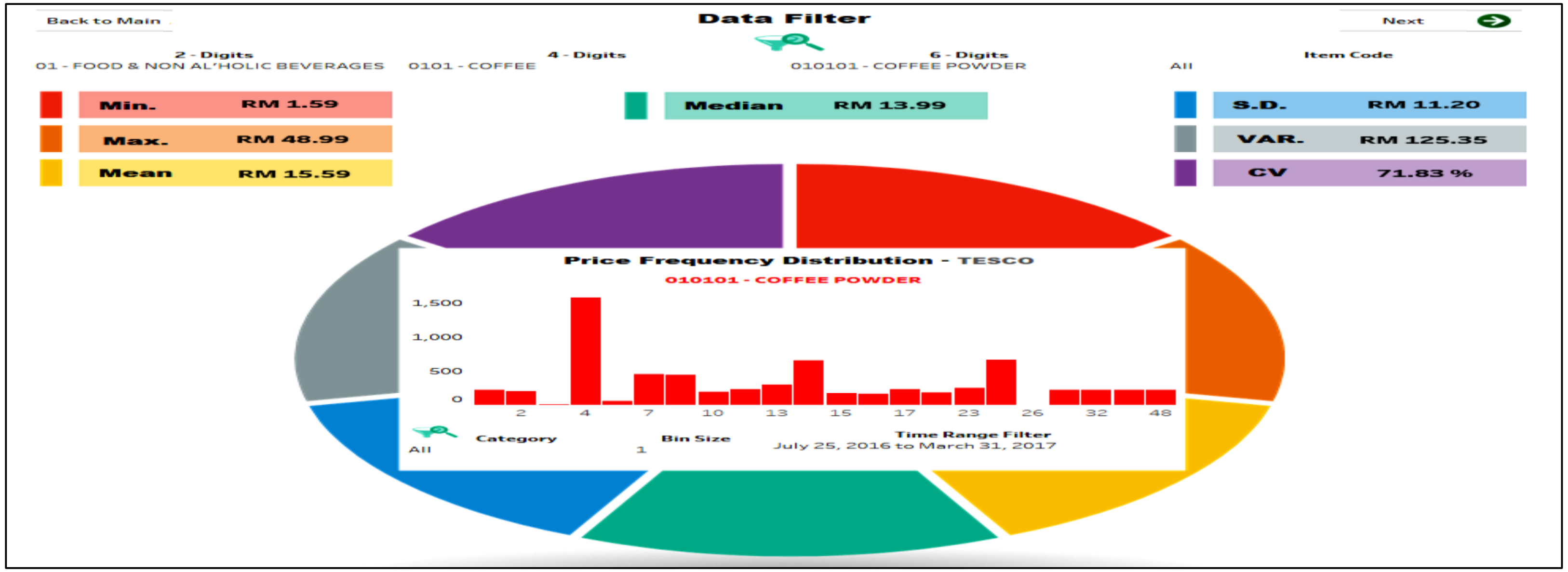
State	CPI Type	Year	Month
SELANGOR	All	2017	All

a)Trend

Trend analysis is used to evaluate data patterns based on linear approaches. There are various time series analysis techniques that can be used, such as ARIMA, Exponential Smoothing, Holt Winter, Linear Trend, Exponential and Level Aggregation.

b) Descriptive

Descriptive analysis, more emphasis on data exploration summaries such as: Mean, Median, Standard Deviation , Variants, Histogram / Skewness, etc.



Data Filter

2 - Digits	4 - Digits	6 - Digits	Item Code
01 - FOOD & NON AL'HOLIC BEVERAGES	0101 - COFFEE	010101 - COFFEE POWDER	All

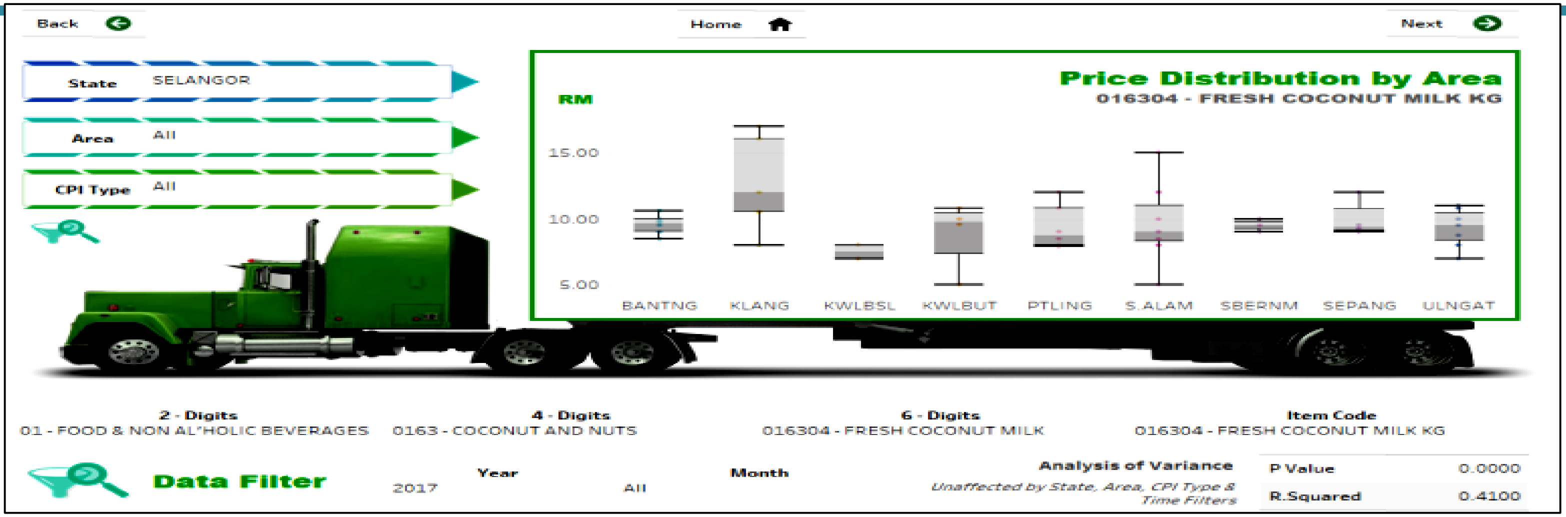
Min.	RM 1.59	
Max.	RM 48.99	
Mean	RM 15.59	
Median	RM 13.99	
S.D.	RM 11.20	
VAR.	RM 125.35	
CV	71.83 %	

Category	Bin Size	Time Range Filter
All	1	July 25, 2016 to March 31, 2017

PRICE INTELLIGENCE ANALYTICS

c) Price Basket Enrichment

Price Basket Enrichment is the process of adding data crawled from online sources into available data from the SIHP-CPI System. This process requires DOSM support because not all prices used in CPI manual calculation are available online.



State SELANGOR

Area All

CPI Type All

Price Distribution by Area
016304 - FRESH COCONUT MILK KG

2 - Digits 01 - FOOD & NON AL'HOLIC BEVERAGES

4 - Digits 0163 - COCONUT AND NUTS

6 - Digits 016304 - FRESH COCONUT MILK

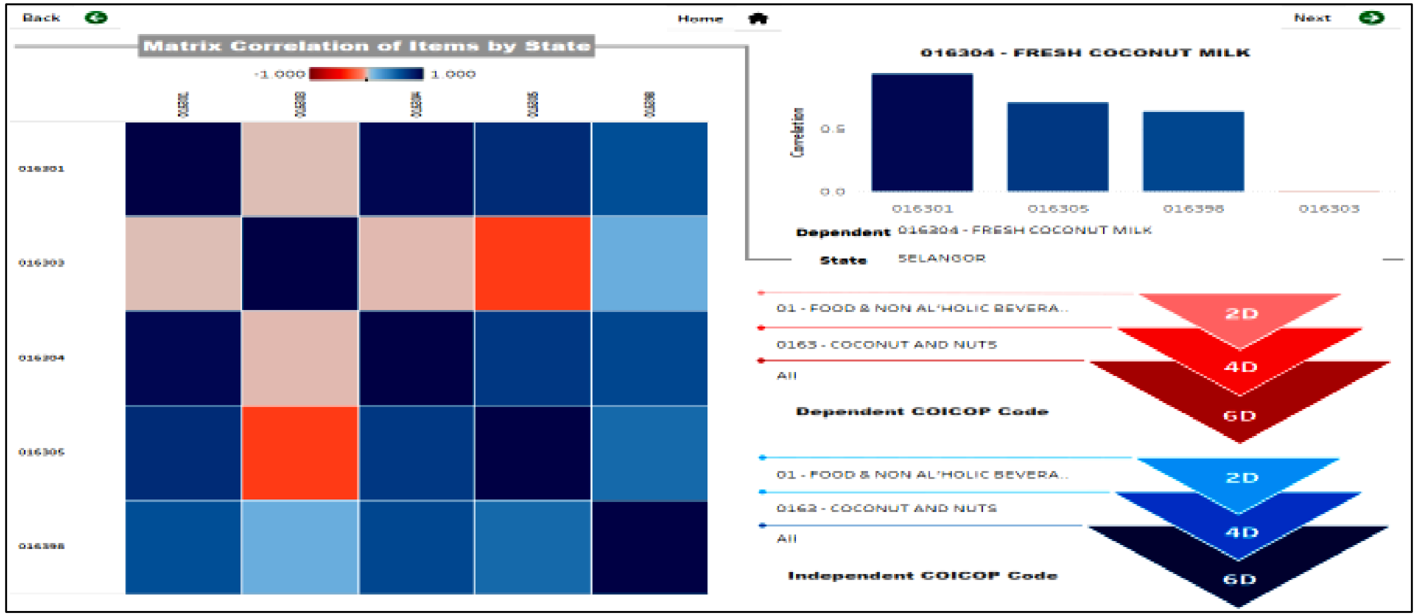
Item Code 016304 - FRESH COCONUT MILK KG

Year 2017

Month All

Analysis of Variance
Unaffected by State, Area, CPI Type & Time Filters

P Value	0.0000
R.Squared	0.4100



Matrix Correlation of Items by State

016304 - FRESH COCONUT MILK

Dependent 016304 - FRESH COCONUT MILK

State SELANGOR

01 - FOOD & NON AL'HOLIC BEVERA..
0163 - COCONUT AND NUTS
All
Dependent COICOP Code

01 - FOOD & NON AL'HOLIC BEVERA..
0163 - COCONUT AND NUTS
All
Independent COICOP Code

d) *Dependency*

Dependency analysis looks for more connections between data. An example is to evaluate variables by performing factor analysis or releasing irrelevant data. Cluster analysis can be carried out to assess data relationships (correlations) and other analysis can be conducted to assess such as Factor Analysis - Cause Analysis, Correlation and Cross Category

APPLICATIONS OF PRICE INTELLIGENCE

During the pandemic Covid-19, online flight ticket prices has been used in the compilation of the CPI.

<u>Item specification (Route)</u>

- ❑ Kedah to Kuala Lumpur
- ❑ Johor Bahru to Kuala Lumpur
- ❑ Kuantan to Kuala Lumpur
- ❑ Penang to Kuala Lumpur
- ❑ Kuala Terengganu to Kuala Lumpur

MAC 2020

Implications of the Movement Control Order (MCO), all the price data collection at the outlet has been suspended. Data crawling has been done during the MCO for 20 main product CPI as below:

- CAP RAMBUTAN HIJAU SST RICE 5% 10KG (RICE)
- JATI BERAS SUPER SPESIAL 10KG (RICE)
- JASMINE RICE SUPER SPECIAL 5% 10KG (RICE)
- JASMINE RICE SUPER SPECIAL TEMPATAN 5KG (RICE)
- AYAM BERSIH (PELBAGAI BAHAGIAN) (CHICKEN)
- IKAN BAWAL HITAM (FISH)
- IKAN CENCARU (FISH)
- IKAN KEMBUNG (FISH)
- KUETIAU BASAH (FLAT RICE NOODLES)
- MEE KUNING BASAH (NOODLES)
- UBI KENTANG (POTATO)
- BAWANG BESAR (ONION)
- CILI KERING KERINTING (DRIED CHILLI)
- KACANG BUNCIS (FRENCH BEAN)
- KUBIS BULAT (TEMPATAN) (CABBAGE)
- LOBAK MERAH (CARROT)
- TOMATO
- SANTAN KELAPA (FRESH COCONUT MILK)
- TELUR AYAM GRED B (HEN'S EGGS GRADE B)
- *MINYAK MASAK (PELBAGAI JENAMA) (COOKING OIL)

Department of Statistics
MALAYSIA

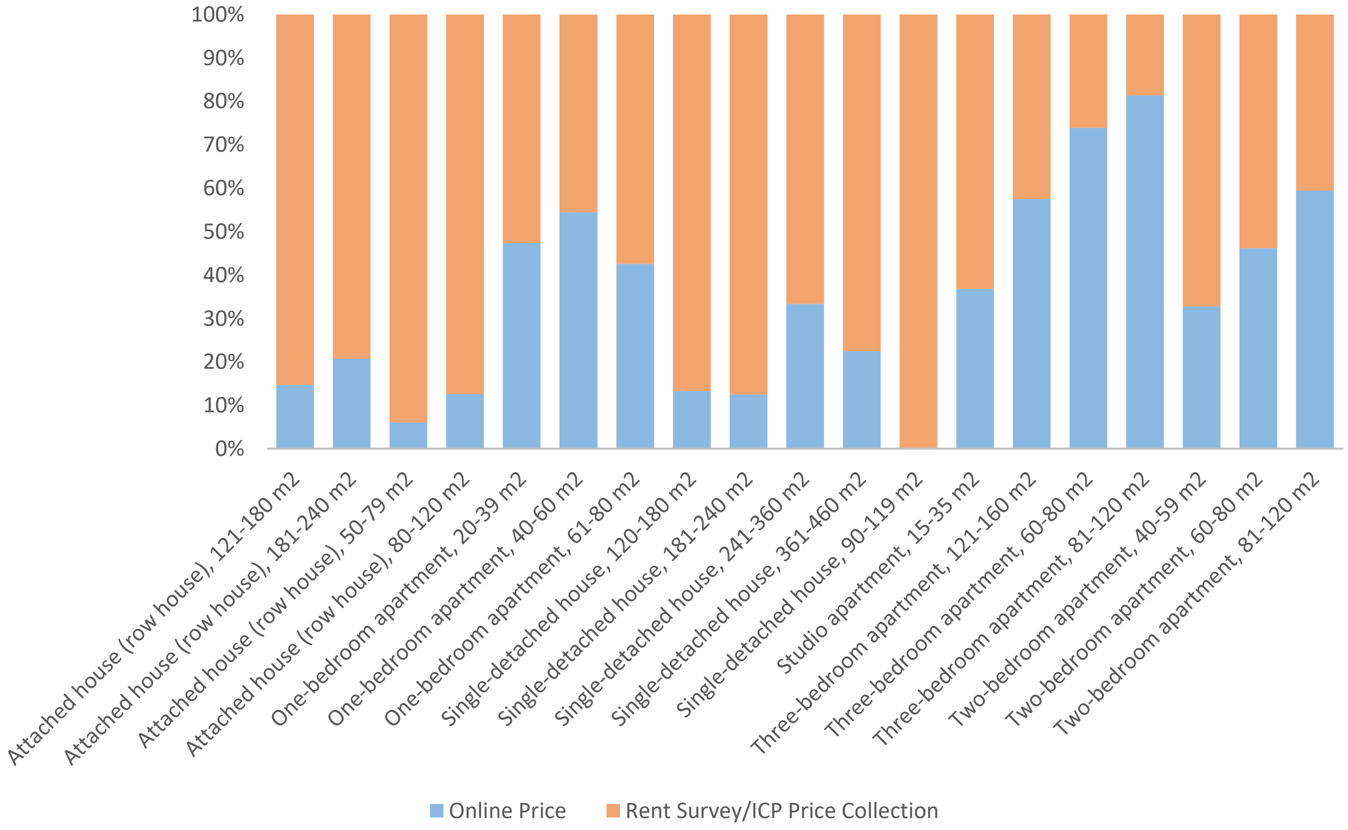
APPLICATIONS OF PRICE INTELLIGENCE

POPULATION & HOUSING CENSUS MALAYSIA 2020
YOUR DATA IS OUR FUTURE

HOUSE RENTAL PRICES FOR INTERNATIONAL COMPARISON PROGRAM (ICP)

45.6% housing rental data for International Comparison Program 2017 (ICP2017) submission were using online price data

RENTAL HOUSING DATA SOURCED USED FOR ICP 2017

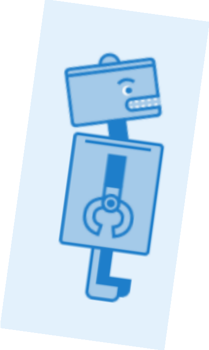


19 types of housing specification with different sizes consisting of

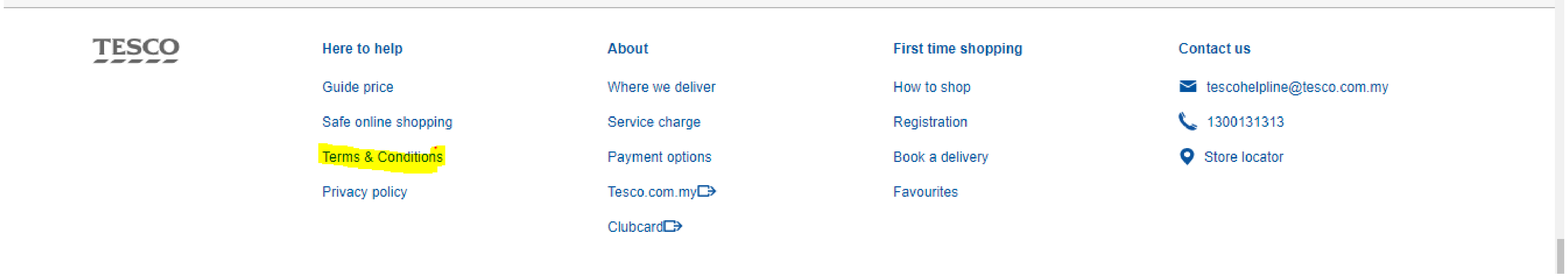
- Single-detached house
- Attached house (row house)
- Studio apartment
- One-bedroom apartment
- Two-bedroom apartment
- Three-bedroom apartment

LIMITATION

- ❑ Need to update crawler
- ❑ Have to build a bunch of crawlers for different sites
- ❑ The structure of websites change frequently
- ❑ Legal issues involved
- ❑ Storage limitation (huge amounts of data)
- ❑ Access and scrape data which is publicly available and avoid trying to crawl data which is private or protected by copyrights and other laws;
- ❑ Always check the website's robots.txt file



LEGALISATION



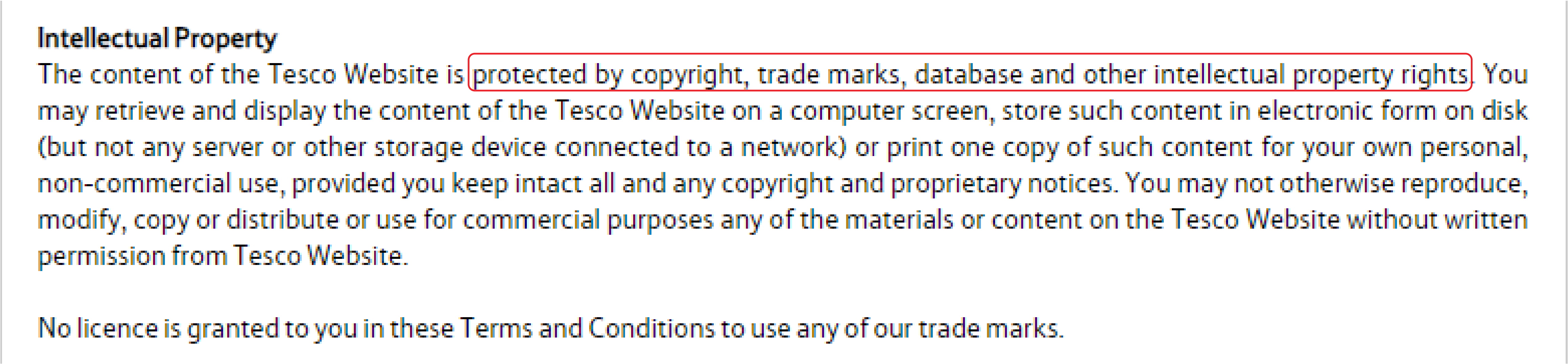
TESCO

Here to help
- Guide price
- Safe online shopping
- Terms & Conditions
- Privacy policy

About
- Where we deliver
- Service charge
- Payment options
- Tesco.com.my
- Clubcard

First time shopping
- How to shop
- Registration
- Book a delivery
- Favourites

Contact us
- ✉ tescohelpline@tesco.com.my
- ☎ 1300131313
- Store locator



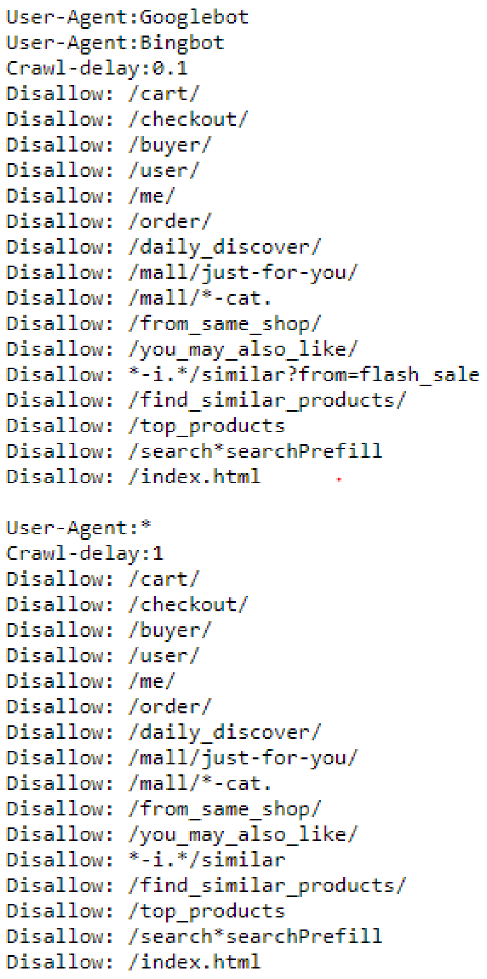
Intellectual Property

The content of the Tesco Website is protected by copyright, trade marks, database and other intellectual property rights. You may retrieve and display the content of the Tesco Website on a computer screen, store such content in electronic form on disk (but not any server or other storage device connected to a network) or print one copy of such content for your own personal, non-commercial use, provided you keep intact all and any copyright and proprietary notices. You may not otherwise reproduce, modify, copy or distribute or use for commercial purposes any of the materials or content on the Tesco Website without written permission from Tesco Website.

No licence is granted to you in these Terms and Conditions to use any of our trade marks.

WHAT IS ROBOTS.TXT?

https://shopee.com.my/robots.txt



```
User-Agent:Googlebot
User-Agent:Bingbot
Crawl-delay:0.1
Disallow: /cart/
Disallow: /checkout/
Disallow: /buyer/
Disallow: /user/
Disallow: /me/
Disallow: /order/
Disallow: /daily_discover/
Disallow: /mall/just-for-you/
Disallow: /mall/*-cat.
Disallow: /from_same_shop/
Disallow: /you_may_also_like/
Disallow: *-i.*/similar?from=flash_sale
Disallow: /find_similar_products/
Disallow: /top_products
Disallow: /search*searchPrefill
Disallow: /index.html

User-Agent:*
Crawl-delay:1
Disallow: /cart/
Disallow: /checkout/
Disallow: /buyer/
Disallow: /user/
Disallow: /me/
Disallow: /order/
Disallow: /daily_discover/
Disallow: /mall/just-for-you/
Disallow: /mall/*-cat.
Disallow: /from_same_shop/
Disallow: /you_may_also_like/
Disallow: *-i.*/similar
Disallow: /find_similar_products/
Disallow: /top_products
Disallow: /search*searchPrefill
Disallow: /index.html
```

A robots.txt file tells search engine crawlers which pages or files the crawler can or can't request from your site. This is used mainly to avoid overloading your site with requests; **it is not a mechanism for keeping a web page out of Google.** To keep a web page out of Google, you should use noindex directives, or password-protect your page.

A robots.txt file is used primarily to manage crawler traffic to your site, and *usually* to keep a file off Google, depending on the file type.

Source:
https://developers.google.com/search/docs/advanced/robots/intro

Department of Statistics
MALAYSIA

RESEARCH PAPER USING WEB SCRAPING (PRICE INTELLIGENCE)

POPULATION & HOUSING CENSUS MALAYSIA 2020
YOUR DATA IS OUR FUTURE

1 Forecasting prices of fish and vegetable using web scraped price micro data

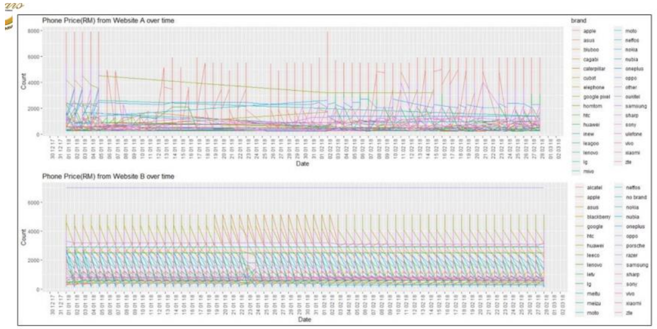
- Mazliana Mustapa, Raja Rajeswari Ponnusamy, Ho Ming Kang

2 Online and Offline Prices: Measuring selected home appliance's products

- Mohd Saiful Husain and Norsyela Muhammad Noor Mathivanan

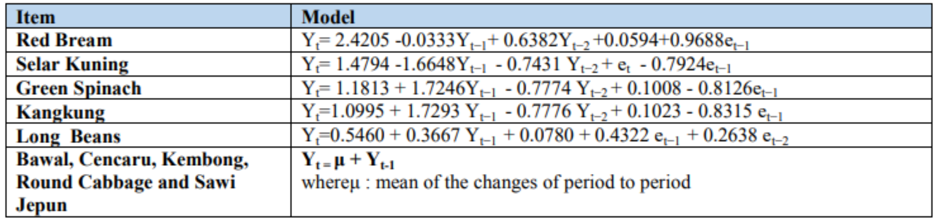
3 Analysis of the Mobile Phones Prices Malaysia using Web Scraped Data

- Nur Hurriyatul Huda Abdullah Sani

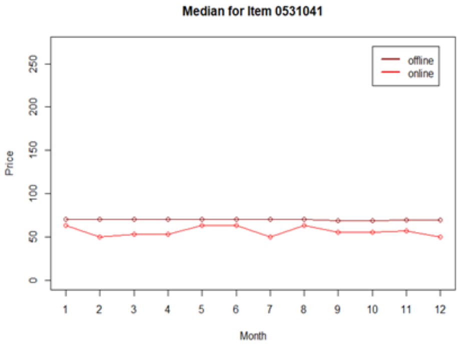

Phone Price(RM) from Website A over time

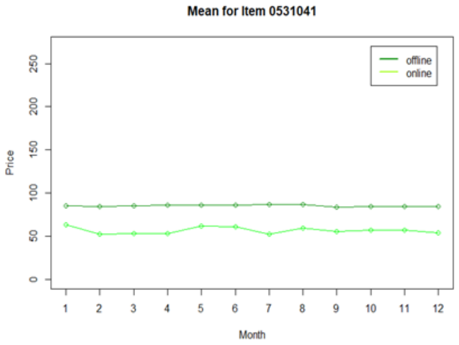
Phone Price(RM) from Website B over time

Table 3: ARIMA best model



Item	Model
Red Bream	$Y_t = 2.4205 - 0.0333Y_{t-1} + 0.6382Y_{t-2} + 0.0594 + 0.9688e_{t-1}$
Selar Kuning	$Y_t = 1.4794 - 1.6648Y_{t-1} - 0.7431\ Y_{t-2} + e_t - 0.7924e_{t-1}$
Green Spinach	$Y_t = 1.1813 + 1.7246Y_{t-1} - 0.7774\ Y_{t-2} + 0.1008 - 0.8126e_{t-1}$
Kangkung	$Y_t = 1.0995 + 1.7293\ Y_{t-1} - 0.7776\ Y_{t-2} + 0.1023 - 0.8315\ e_{t-1}$
Long Beans	$Y_t = 0.5460 + 0.3667\ Y_{t-1} + 0.0780 + 0.4322\ e_{t-1} + 0.2638\ e_{t-2}$
Bawal, Cencaru, Kembong, Round Cabbage and Sawi Jepun	$Y_{t} = \mu + Y_{t-1}$ whereµ : mean of the changes of period to period


Median for Item 0531041


Mean for Item 0531041

CONTACT INFO

https://statsbda.dosm.gov.my/

Mazliana Mustapa
Core Team Big Data Analytics
Department of Statistics Malaysia

Team Price Intelligence
mazliana@dosm.gov.my
ridhuan@dosm.gov.my
noradilah.adnan@dosm.gov.my

THANK YOU

BANCI MALAYSIA

POPULATION & HOUSING CENSUS MALAYSIA 2020

YOUR DATA IS OUR FUTURE

#MyCensus2020

Make Sure You're COUNTED

For More Info : www.dosm.gov.my
www.mycensus.gov.my

@StatsMalaysia
@MyCensus2020

MYSTATS DAY 20 OCT

2016-2030